We thank the reviewers for their constructive feedback. R1 and R2 appreciated the novelty of our unified framework that solves both feature imputation and label prediction aspects for the missing data problem, and all reviewers agreed that our solution is well motivated. Reviewers pointed out several main concerns, which we summarize and answer below:

1 Novelty of the GNN method (R3 R4). We thank R3 and R4 for noticing that the core GNN components have been separately used in other applications, and the formulation as a bipartite graph has been used for matrix completion task. However, we emphasize that our main contribution is *not the particular GNN model but the graph-based framework*. We show that a seemingly unrelated missing data problem (imputation and learning subsequent tasks) can naturally be solved with graphs and we propose the first graph-based solution to it, which is acknowledged by R1 and R2. Nevertheless, we agree with R3 and R4 that the justification of the adopted method can be further improved and we will do that in the final version of the paper. **New results to justify our model.** In addition to ablation study in Section 4.6, we further justify our model by showing how different types of aggregation, *i.e.*, MEAN(GraphSAGE-mean), SUM(GIN), MAX(GraphSAGE-pool) affect the performance. These new results justify our selection of GraphSAGE-style pooling in our architecture. **Justifying the new results.** While SUM is theoretically most expressive, in our setting the degree of a specific node is determined by the number of missing values which is random and unrelated to the missing data task; in contrast, the MEAN and MAX aggregators are not affected by this inherent randomness of node degree. We will add these studies to justify our framework.

	concrete	energy	housing	kin8nm	naval	power	protein	wine	yacht
Sum	0.094	0.143	0.078	0.277	0.024	0.134	0.040	0.069	0.154
Max	**0.088**	0.142	**0.074**	0.252	**0.006**	**0.102**	**0.024**	**0.063**	0.153
Mean	0.090	**0.136**	0.075	**0.249**	0.008	**0.102**	0.027	**0.063**	**0.151**

2 Comparing with more baselines (R1 R2 R3). We thank reviewers for pointing out additional baselines to compare, in addition to the 7 baselines that we have already examined (lines 169–181). **New baseline results.** Summarizing reviewer's suggestions, we additionally compared our methods with 3 representative baselines on the feature imputation task: (1) low-rank matrix completion method for mixed-type data (missMDA), (2) deep latent variable models and autoencoders (MIWAE), (3) GNN-based baselines (GC-MC (Berg et al 2017) [1], IGMC (Zhang et al 2019) [2]). **Discussion.** Note that *all the new baselines are only applicable to the feature imputation task* not the label prediction task, and *the GNN-based baselines are only applicable to discrete-value data*, as they explicitly use different weights for each distinct value, and thus cannot apply to mix-typed data; GRAPE *has none of these limitations*. We find that our GRAPE framework can consistently outperform these new baselines in all the datasets. In 3 discrete-value datasets, IGMC [2] has advantages as it is specifically designed for discrete-value data, while our model can still outperform GC-MC [1]. We will include these new results and make sure to cite all the papers suggested in the reviews.

	concrete	energy	housing	kin8nm	naval	power	protein	wine	yacht
missMDA	0.190	0.225	0.142	0.285	0.038	0.215	0.068	0.090	0.226
MIWAE	0.156	0.153	0.098	0.262	0.020	0.117	0.042	0.087	0.224
Ours	**0.090**	**0.136**	**0.075**	**0.249**	**0.008**	**0.102**	**0.027**	**0.063**	**0.151**

	Flixster	Douban	Yahoo
GC-MC [1]	0.917	0.734	20.5
IGMC [2]	**0.872**	**0.721**	**19.1**
Ours	0.899	0.733	19.4

3 Scalability (R2 R3 R4). We thank reviewers for asking to discuss the scalability of our method. We pick UCI as they are widely-used datasets for benchmarking imputation methods, with *both discrete and continuous features*. Our GRAPE *framework does not suffer from scalability issues*: the core of our method is a GNN, which has been successfully applied to graphs with billions of edge; while R3 pointed out that a naive one-hot scheme will not work when the number of features is prohibitively large, we can use an embedding matrix to encode each feature, which has successfully handled millions of tokens in NLP. We agree with reviewers that experimenting on datasets with more features can justify the scalability. We provide new results on large-scale benchmarks (Flixster: 2956 features, Douban: 3000 features, Yahoo: 1363 features) shown in the table above. Note that these datasets *only have discrete features*, and therefore they were not used in our manuscript. We will include these new discussions and results in the revised paper.

4 Discussions of related work (R1). We thank R1 for pointing out that our discussions for the related work can be improved. We intended to give an overview for common limitations in lines 23–31 by starting with "often exhibit notable shortcomings", but we do agree that this summary should be more rigorous. **Clarifications.** (1) We will cite these matrix completion models designed for both discrete and continuous variables, and for online learning. (2) We will change to "imputation approaches based on deep generative models do not *explicitly* use feature values from other observations". (3) After formulating the missing data problem as a bipartite graph over feature and observation nodes, the proposed augmented node feature initialization is natural: feature nodes are not permutation invariant so one-hot tokens are used; observation nodes are permutation invariant so constant node features are used – we have provided a thorough discussion in lines 130–149. Overall, we will include these discussions in the revised paper.

5 Clarifications. **Q:** (R2)"Applicable to vision data" **A:** Yes, *e.g.*, our model can be jointly used with CNNs and trained end-to-end. **Q:** (R2 R4)"Theoretical guarantees" **A:** The success of our framework is supported by the universal approximation capabilities of GNNs on graph structured data. **Q:** (R4) "Random missing data assumption" **A:** This is the most common evaluation regime used in missing data papers. Our model can apply to other missing data scenarios. **Q:** (R4) "Using labels in feature imputation" **A:** We agree that label information could be helpful in feature imputation but we do not use it in our experiments, since the common evaluation for feature imputation tasks does not use labels.